Journal of Multivariate Analysis 103 (2012) 48–57

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Testing for positive evidence of equally likely outcomes

Jesse Frey

Department of Mathematical Sciences, Villanova University, Villanova, PA 19085, United States

ARTICLE INFO

Article history:
Received 18 October 2010
Available online 21 June 2011

AMS subject classifications:
62H15
62F03
62F25

Keywords:
Digits of π
Equivalence testing
Intersection–union testing
Multinomial distribution
Roulette
Testing uniformity

ABSTRACT

Goodness-of-fit tests allow one to conclude that k possible outcomes are not equally likely. In this paper, we develop an exact equivalence test that allows one to conclude that k possible outcomes are *approximately* equally likely. We show that the power properties of the test compare favorably to those of possible alternative tests, and we develop an associated simultaneous confidence interval procedure. We apply the test to data sets on the digits of π, winning roulette numbers, and winning numbers from the Pennsylvania Lottery.

1. Introduction

When evaluating gambling devices, lotteries, and random number generators, one hopes to assess whether certain sets of outcomes are equally likely or not. Goodness-of-fit tests allow one to obtain evidence that the outcomes are not equally likely. However, if one wishes to establish that the outcomes are approximately equally likely, one must use a different approach.

Suppose that each run of a random process leads to exactly one of k possible outcomes. If N_i is the number of times the ith outcome occurs in N independent runs, then $(N_1, \ldots, N_k) \sim \text{Multi}(N, \mathbf{p})$, where $\mathbf{p} \equiv (p_1, \ldots, p_k)$ is the vector of outcome probabilities. Given a candidate vector $\mathbf{p_0} \equiv (p_{10}, \ldots, p_{k0})$, it is natural to test $H_0 : \mathbf{p} = \mathbf{p_0}$ against $H_1 : \mathbf{p} \neq \mathbf{p_0}$. A variety of tests for these goodness-of-fit hypotheses are available, including the chi-squared test, the likelihood ratio test, and other tests that were studied by Cressie and Read [4]. For the case of equally likely outcomes, which was recently considered by Spencer [17], the candidate vector is $\mathbf{p_0} = (1/k, \ldots, 1/k)$. Rejecting $H_0 : \mathbf{p} = (1/k, \ldots, 1/k)$ in favor of $H_1 : \mathbf{p} \neq (1/k, \ldots, 1/k)$ allows us to conclude that the outcomes are not equally likely. However, to establish that the outcomes are approximately equally likely, we must replace the goodness-of-fit test with an equivalence test.

Equivalence testing involves reversing the null and alternative hypotheses used in goodness-of-fit testing, while also introducing a zone of indifference (equivalence region) around the original point null hypothesis. For example, the standard two-sample problem involves testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, where μ_1 and μ_2 are the means for two separate populations. The corresponding equivalence testing hypotheses are $H_0 : |\mu_1 - \mu_2| \geq \Delta$ and $H_1 : |\mu_1 - \mu_2| < \Delta$, where Δ is a user-chosen distance small enough that differences less than Δ have little importance. With these new hypotheses, rejecting H_0 allows us to conclude that μ_1 and μ_2 are approximately the same. Equivalence tests were first developed by Lehmann [15] and Bondy [3], but the work on bioequivalence testing by Westlake [19] and others brought them to greater prominence. Many other applications of equivalence testing have also been developed. For example, Dixon and Pechmann [7] developed an equivalence test that allows one to conclude that a linear trend is of negligible strength, and Robinson and

E-mail address: jesse.frey@villanova.edu.

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2011.06.006

Froese [16] used equivalence testing in model validation. Additional applications of equivalence testing were discussed by Wellek [18].

Equivalence testing using multinomial data is not new. Wellek [18, Section 8.1] developed a test of $H_0 : d_{Euc}(\mathbf{p}, \mathbf{p_0}) \geq \Delta$ against $H_1 : d_{Euc}(\mathbf{p}, \mathbf{p_0}) < \Delta$, where $d_{Euc}(\mathbf{p_1}, \mathbf{p_2})$ is the Euclidean distance between $\mathbf{p_1}$ and $\mathbf{p_2}$. However, the Wellek [18] test is valid only asymptotically, and the notion of equivalence used by Wellek [18] does not have clear implications for the individual outcome probabilities. Frey [10] developed an exact test of $H_0 : d_{Max}(\mathbf{p_1}, \mathbf{p_2}) \geq \Delta$ against $H_1 : d_{Max}(\mathbf{p_1}, \mathbf{p_2}) < \Delta$, where $d_{Max}(\mathbf{p_1}, \mathbf{p_2})$ is defined by $d_{Max}(\mathbf{p_1}, \mathbf{p_2}) \equiv \max_{i=1,\ldots,k-1} | \sum_{j=1}^{i}(p_{j1} - p_{j2})|$. Thus, the distance $d_{Max}(\mathbf{p_1}, \mathbf{p_2})$ is obtained by computing all partial sums of the k-vectors $\mathbf{p_1}$ and $\mathbf{p_2}$ and then finding the maximum absolute difference between corresponding partial sums. The Frey [10] test is exact, but the distance measure $d_{Max}(\cdot, \cdot)$ is not invariant under permutation of the outcome labels $1, \ldots, k$. Thus, neither of these existing equivalence tests is ideal for obtaining positive evidence of approximately equally likely outcomes.

In this paper, we define equivalence in a manner that is invariant under permutation of the outcome labels $1, \ldots, k$. Rather than test $H_0 : \mathbf{p} = (1/k, \ldots, 1/k)$ against $H_1 : \mathbf{p} \neq (1/k, \ldots, 1/k)$, we test $H_0 : \mathbf{p} \notin (a, b)^k$ against $H_1 : \mathbf{p} \in (a, b)^k$, where a and b are user-chosen bounds satisfying $0 \leq a < 1/k < b \leq 1$. With the hypotheses defined in this way, rejection of H_0 allows us to conclude that $p_i \in (a, b)$ for $i = 1, \ldots, k$. Thus, if the interval (a, b) has been chosen to be sufficiently narrow, then we are able to conclude that $p_1, \ldots, p_k$ are approximately equal. In choosing the equivalence region, it is tempting to consider letting the equivalence region include only the single point $(1/k, \ldots, 1/k)$. However, in this case, the equivalence region has no interior. As a result, the power of the test cannot exceed the chosen α level. Thus, in what follows, we use an equivalence region $(a, b)^k$ for a and b satisfying $0 \leq a < 1/k < b \leq 1$.

We describe a conservative intersection–union test of $H_0 : \mathbf{p} \notin (a, b)^k$ against $H_1 : \mathbf{p} \in (a, b)^k$ in Section 2, and we also show how to improve the intersection–union test by using an algorithm for computing rectangular multinomial probabilities. We invert the improved test to obtain simultaneous confidence intervals for $p_1, \ldots, p_k$ in Section 3, and we compare the power of the test to that of possible alternative tests for the same hypotheses in Section 4. In Section 5, we apply the test and the confidence interval procedure to data sets on the digits of π, winning roulette numbers, and winning numbers from the Pennsylvania Lottery. We discuss an alternate method for choosing the equivalence region in Section 6, and we give conclusions in Section 7.

2. The intersection–union test and an improvement

Suppose that values a, b satisfying $0 < a < 1/k < b < 1$ are given. We wish to test $H_0 : \mathbf{p} \notin (a, b)^k$ against $H_1 : \mathbf{p} \in (a, b)^k$ using data $(N_1, \ldots, N_k) \sim \text{Multi}(N, \mathbf{p})$. If $\mathbf{p} \notin (a, b)^k$, then either $p_i \leq a$ for some i or $p_i \geq b$ for some i. Thus, if we define $L_i \equiv \{\mathbf{p} : p_i \leq a\}$ and $U_i \equiv \{\mathbf{p} : p_i \geq b\}$ for $i = 1, \ldots, k$, then we may write the null hypothesis set $\{\mathbf{p} : \mathbf{p} \notin (a, b)^k\}$ as $(\bigcup_i L_i) \bigcup (\bigcup_i U_i)$. The theory of intersection–union testing [2] then tells us that if we test each of the hypotheses $H_{0i}^- : p_i \leq a$ and $H_{0i}^+ : p_i \geq b$ for $i = 1, \ldots, k$ at level α, then the test of H_0 against H_1 that consists of rejecting H_0 if and only if all $2k$ null hypotheses are rejected is also a level-α test.

Each hypothesis $H_{0i}^- : p_i \leq a$ or $H_{0i}^+ : p_i \geq b$ can be tested in a natural way by using the fact that $N_i \sim \text{Bin}(N, p_i)$. The p-value for testing H_{0i}^- is $P(X \geq N_i \mid X \sim \text{Bin}(N, a))$, and the p-value for testing H_{0i}^+ is $P(X \leq N_i \mid X \sim \text{Bin}(N, b))$. Thus, the appropriate critical values for tests of the two types are c_l (for tests of H_{0i}^-) and c_u (for tests of H_{0i}^+), where c_l is the smallest integer such that $P(X \geq c_l \mid X \sim \text{Bin}(N, a)) \leq \alpha$ and c_u is the largest integer such that $P(X \leq c_u \mid X \sim \text{Bin}(N, b)) \leq \alpha$. When c_l and c_u are given, the intersection–union test consists of rejecting $H_0 : \mathbf{p} \notin (a, b)^k$ in favor of $H_1 : \mathbf{p} \in (a, b)^k$ if and only if $c_l \leq n_i \leq c_u$ for $i = 1, \ldots, k$.

The intersection–union test may also be carried out by computing an overall p-value. This p-value is obtained as $p_{IU} = \max\{P(X \leq n_{max} \mid X \sim \text{Bin}(N, b)), P(X \geq n_{min} \mid X \sim \text{Bin}(N, a))\}$, where n_{min} and n_{max} are the smallest and largest of the observed counts $\{n_1, \ldots, n_k\}$. If $p_{IU} \leq \alpha$, then we reject H_0 at level α. Otherwise, we retain the null hypothesis $H_0 : \mathbf{p} \notin (a, b)^k$.

The intersection–union test just described provides guaranteed control of the Type I error rate, but it can be highly conservative when the sample size is small. Indeed, for small N, it is not unusual to have $c_l > c_u$ so that the test has true size 0. However, we show in what follows that for any choice of c_l and c_u, the exact size of the test can be computed. As a result, it is possible to adjust the level at which the $2k$ hypotheses $H_{01}^-, \ldots, H_{0k}^-, H_{01}^+, \ldots, H_{0k}^+$ are tested in such a way that the size of the overall test comes closer to the desired level α. The key result needed here is Theorem 1, which follows from two lemmas that we prove now.

Lemma 1. *Let N be a fixed integer, and let c be an integer satisfying $0 \leq c \leq N/2$. If*

$$F(p) \equiv P(c \leq X \leq N - c \mid X \sim Bin(N, p)),$$

then $F(p)$ is non-decreasing on $[0, 1/2]$ and non-increasing on $[1/2, 1]$.

Proof. If $c = 0$, then $F(p) = 1$ for all $p \in [0, 1]$, and the claim holds. Suppose now that $c > 0$. Using a familiar connection between the binomial distribution and the beta distribution ([6], p. 160), we may write

$$F(p) = P(c \leq X \leq N - c \mid X \sim \text{Bin}(N, p)) = I_p(c, N - c + 1) - I_p(N - c + 1, c)$$

$$= \int_0^p \frac{N!}{(c-1)!(N-c)!} y^{c-1}(1-y)^{N-c}\,dy - \int_0^p \frac{N!}{(N-c)!(c-1)!} y^{N-c}(1-y)^{c-1}\,dy,$$

where $I_p(\alpha, \beta)$ is the incomplete beta function. Differentiating, we have that

$$\begin{aligned}\frac{dF}{dp} &= \frac{N!}{(c-1)!(N-c)!}\{p^{c-1}(1-p)^{N-c} - p^{N-c}(1-p)^{c-1}\} \\ &= \frac{N!}{(c-1)!(N-c)!} p^{c-1}(1-p)^{c-1}\{(1-p)^{N-2c+1} - p^{N-2c+1}\}.\end{aligned}$$

The factor $p^{c-1}(1-p)^{c-1}$ is always positive, and the factor $(1-p)^{N-2c+1} - p^{N-2c+1}$ is strictly decreasing on the interval $[0, 1]$. Moreover, $(1-p)^{N-2c+1} - p^{N-2c+1}$ is 0 when $p = 1/2$. Thus, $F(p)$ is strictly increasing on $[0, 1/2]$ and strictly decreasing on $[1/2, 1]$. This completes the proof. □

Lemma 2. *Let c_l and c_u be fixed integers satisfying $0 \le c_l \le c_u \le N$, and define $G(\mathbf{p})$ by*

$$G(\mathbf{p}) \equiv P(c_l \le N_i \le c_u \forall i \mid (N_1, \ldots, N_k) \sim \mathit{Multi}(N, \mathbf{p})). \tag{1}$$

Let $\mathbf{p} \equiv (p_1, \ldots, p_k)$ be an arbitrary probability vector, let $\lambda \in [0, 1]$, and let $i \ne j$ be two values from the set $\{1, \ldots, k\}$. If we define a new probability vector $\mathbf{p}'$ by setting $p_i' = \lambda p_i + (1-\lambda)p_j$, $p_j' = \lambda p_j + (1-\lambda)p_i$, and $p_l' = p_l$ for $l \ne i, j$, then $G(\mathbf{p}') \ge G(\mathbf{p})$.

Proof. Note that $G(\mathbf{p})$ is invariant under permutation of the labels $1, \ldots, k$. Thus, without loss of generality, we may take $i = 1$ and $j = 2$. We now use a conditioning argument. Let $n_3, \ldots, n_k$ be given values for $N_3, \ldots, N_k$. The conditional distribution of N_1 is then $\mathrm{Bin}(t, \frac{p_1}{p_1+p_2})$, where $t = N - (n_3 + \cdots + n_k)$, and $N_2 = t - N_1$. Let $c, c+1, \ldots, d$ be the list of values for N_1 that would give us $c_l \le N_i \le c_u$ for $i = 1, 2$. It then follows by symmetry that $d = t - c$. Thus, the conditional probability that $c_l \le N_i \le c_u$ for $i = 1, 2$ is $P(c \le X \le t - c \mid X \sim \mathrm{Bin}(t, p_1/(p_1+p_2)))$. By Lemma 1, such probabilities either increase or stay the same when $p_1/(p_1+p_2)$ is made closer to $1/2$. Thus, replacing $\mathbf{p}$ with $\mathbf{p}'$ can only increase the conditional probability that $c_l \le N_i \le c_u$ for $i = 1, 2$.

Let S be the set of all possible values for the vector $(N_3, \ldots, N_k)$. Since the distribution of $(N_3, \ldots, N_k)$ is unchanged when we replace $\mathbf{p}$ with $\mathbf{p}'$, it follows that

$$\begin{aligned}G(\mathbf{p}') &= \sum_{(n_3,\ldots,n_k)\in S} \{P(N_3 = n_3, \ldots, N_k = n_k) I(c_l \le n_i \le c_u \text{ for } i = 3, \ldots, k) \\ &\quad \cdot P(c_l \le N_i \le c_u \text{ for } i = 1, 2 \mid n_3, \ldots, n_k, \mathbf{p}')\} \\ &\ge \sum_{(n_3,\ldots,n_k)\in S} \{P(N_3 = n_3, \ldots, N_k = n_k) I(c_l \le n_i \le c_u \text{ for } i = 3, \ldots, k) \\ &\quad \cdot P(c_l \le N_i \le c_u \text{ for } i = 1, 2 \mid n_3, \ldots, n_k, \mathbf{p})\} = G(\mathbf{p}),\end{aligned}$$

where $I(A)$ is 1 if A holds and 0 otherwise. This completes the proof. □

Theorem 1. *The true size of the test of $H_0 : \mathbf{p} \notin (a, b)^k$ against $H_1 : \mathbf{p} \in (a, b)^k$ with critical region $\{(n_1, \ldots, n_k) : c_l \le n_i \le c_u$ for $i = 1, \ldots, k\}$ is achieved either at the point $\mathbf{p}_{a,k} \equiv (a, (1-a)/(k-1), \ldots, (1-a)/(k-1))$ or at the point $\mathbf{p}_{b,k} \equiv (b, (1-b)/(k-1), \ldots, (1-b)/(k-1))$.*

Proof. We noted at the beginning of this section that the null hypothesis region may be written as $\left(\bigcup_i L_i\right) \bigcup \left(\bigcup_i U_i\right)$, where $L_i = \{\mathbf{p} : p_i \le a\}$ and $U_i = \{\mathbf{p} : p_i \ge b\}$ for $i = 1, \ldots, k$. The true size of the test is $\sup_{\mathbf{p}\notin(a,b)^k} G(\mathbf{p})$, where $G(\mathbf{p})$ is as defined in (1). By Lemma 2, moving any two elements of the probability vector closer together either increases $G(\mathbf{p})$ or keeps it the same. Thus, for $\mathbf{p} \in L_1$, $G(\mathbf{p})$ must be maximized at $\mathbf{p}_{a,k}$, and for $\mathbf{p} \in U_1$, $G(\mathbf{p})$ must be maximized at $\mathbf{p}_{b,k}$. Noting that $\sup_{\mathbf{p}\in L_i} G(\mathbf{p})$ and $\sup_{\mathbf{p}\in U_i} G(\mathbf{p})$ are independent of i then completes the proof. □

Theorem 1 allows us to find the exact size α_{IU} for the intersection–union test that consists of rejecting H_0 only when all of the $2k$ hypotheses $H_{01}^-, \ldots, H_{0k}^-, H_{01}^+, \ldots, H_{0k}^+$ are rejected at level α. Specifically, the theorem tells us that $\alpha_{IU} = \max\{G(\mathbf{p}_{a,k}), G(\mathbf{p}_{b,k})\}$. Values for $G(\mathbf{p})$ can be computed by writing $G(\mathbf{p})$ as the rectangular multinomial probability given in (1) and then applying the algorithm of Frey [11].

The true size α_{IU} for the intersection–union test necessarily satisfies $\alpha_{IU} \le \alpha$, and often α_{IU} is much smaller than α. To obtain a more powerful test, we simply increase the level at which the $2k$ hypotheses $H_{01}^-, \ldots, H_{0k}^-, H_{01}^+, \ldots, H_{0k}^+$ are tested. Specifically, we find a value α' such that the test of H_0 against H_1 that consists of testing each hypothesis $H_{01}^-, \ldots, H_{0k}^-, H_{01}^+, \ldots, H_{0k}^+$ at level α' and then rejecting if and only if each of the $2k$ hypotheses is rejected has a size α_{new} that comes as close as possible to α without exceeding α. Due to discreteness, the choice of α' is not unique, but the achieved level α_{new} is unique. Since α_{new} is a non-decreasing function of α', it is easy to find an appropriate α' by using a root-finding algorithm such as bisection (see [1]).

Table 1
Critical regions and true sizes for level-0.05 intersection–union tests and level-0.05 improved tests when $k = 5$, $\alpha = 0.05$, and $(a, b) = (0.1, 0.3)$.

N	Intersection–union test		Improved test	
	Critical region	Size	Critical region	Size
20	Empty	.000	$n_i \in [3, 4]\ \forall i$	.002
40	Empty	.000	$n_i \in [6, 9]\ \forall i$	.026
60	Empty	.000	$n_i \in [9, 14]\ \forall i$	.043
80	$n_i = 16\ \forall i$	.000	$n_i \in [12, 18]\ \forall i$	.020
100	$n_i \in [16, 22]\ \forall i$	.008	$n_i \in [15, 23]\ \forall i$	.027
120	$n_i \in [19, 27]\ \forall i$	.013	$n_i \in [17, 28]\ \forall i$	.034
140	$n_i \in [21, 32]\ \forall i$	.017	$n_i \in [20, 33]\ \forall i$	.034
160	$n_i \in [23, 38]\ \forall i$	.033	$n_i \in [23, 38]\ \forall i$	.033
180	$n_i \in [26, 43]\ \forall i$	.031	$n_i \in [25, 44]\ \forall i$	.049
200	$n_i \in [28, 48]\ \forall i$	.029	$n_i \in [27, 49]\ \forall i$	.045

The improved test may also be conducted by computing an overall p-value. One first computes the p-values for testing each of the $2k$ hypotheses $H_{01}^-, \ldots, H_{0k}^-, H_{01}^+, \ldots, H_{0k}^+$. One then computes the maximum p_{IU}, which is the p-value for the intersection–union test. The p-value for the improved test is then given by $p_{new} = \max\{P(P_{IU} \le p_{IU} \mid \mathbf{p} = \mathbf{p}_{a,k}), P(P_{IU} \le p_{IU} \mid \mathbf{p} = \mathbf{p}_{b,k})\}$, where P_{IU} is the p-value for the intersection–union test as a random variable, p_{IU} is the observed p-value, and $\mathbf{p}_{a,k}$ and $\mathbf{p}_{b,k}$ are the probability vectors defined in Theorem 1.

Table 1 illustrates the gains available from using the new test rather than the intersection–union test. The table shows the critical region and the size for each test when $k = 5$, $\alpha = 0.05$, $(a, b) = (0.1, 0.3)$, and $N = 20, 40, \ldots, 200$. We see from the table that, while the intersection–union test has size 0 for $N = 20$, $N = 40$, and $N = 60$, the improved test has a nonzero size in all tabled scenarios. We also see that the critical region for the improved test is strictly larger than that of the intersection–union test for all of the tabled scenarios except the $N = 160$ case.

Thus far in this section, we have presented our results for the case where $0 < a < b < 1$. However, the same ideas may be used to test $H_0^- : \mathbf{p} \notin (a, 1]^k$ against $H_1^- : \mathbf{p} \in (a, 1]^k$ or $H_0^+ : \mathbf{p} \notin [0, b)^k$ against $H_1^+ : \mathbf{p} \in [0, b)^k$. The necessary modification is that instead of testing all $2k$ hypotheses, one tests only the k hypotheses $H_{01}^-, \ldots, H_{0k}^-$ (to test H_0^-) or the k hypotheses $H_{01}^+, \ldots, H_{0k}^+$ (to test H_0^+). The intersection–union test proceeds by looking at the maximum p_{IU} of the k p-values, and the improved test proceeds by looking at the maximum chance of seeing a value p_{IU} as small or smaller than the one actually observed. The analog of Theorem 1 for these one-sided cases is that the size of the test is achieved either at the point $\mathbf{p}_{a,k}$ (when testing H_0^-) or at the point $\mathbf{p}_{b,k}$ (when testing H_0^+). The hypotheses H_0^+ and H_1^+ are almost exactly the reverse of those considered by Ethier [8], who developed a test that allows one to conclude that there are numbers on a roulette wheel that favor the gambler. However, because the hypotheses are reversed, the probability vector at which the size of Ethier's test is achieved is different from $\mathbf{p}_{b,k}$.

3. Simultaneous confidence intervals

Confidence sets are typically obtained by inverting hypothesis tests. To obtain simultaneous confidence intervals for $p_1, \ldots, p_k$ in this setting, it is helpful to first reduce the number of parameters from two (a and b) to one. Thus, given a value $\Delta > 0$, we define $p_1, \ldots, p_k$ to be equivalent if $|p_i - 1/k| < \Delta$ for all i. Testing equivalence then requires testing $H_{0,\Delta} : \max_i |p_i - 1/k| \ge \Delta$ against $H_{1,\Delta} : \max_i |p_i - 1/k| < \Delta$, and this testing can be done either by setting $(a, b) = (1/k - \Delta, 1/k + \Delta)$ and testing $H_0 : \mathbf{p} \notin (a, b)^k$ against $H_1 : \mathbf{p} \in (a, b)^k$ or, if $\Delta > 1/k$, by setting $b = 1/k + \Delta$ and testing $H_0^+ : \mathbf{p} \notin [0, b)^k$ against $H_1^+ : \mathbf{p} \in [0, b)^k$. If we reject $H_{0,\Delta}$, then we are able to conclude that each p_i is contained in the interval $(1/k - \Delta, 1/k + \Delta)$. Thus, if $\Delta_\alpha \equiv \sup\{\Delta : H_{0,\Delta}$ is not rejected at level $\alpha\}$, then $(1/k - \Delta_\alpha, 1/k + \Delta_\alpha)$ is a simultaneous $100(1 - \alpha)\%$ confidence interval for $p_1, \ldots, p_k$.

If the p-value for testing $H_{0,\Delta}$ were a continuous, decreasing function of Δ, then we could find Δ_α by using a root-finding algorithm like bisection. However, the p-value is not a continuous function of Δ. Instead, it is continuous and decreasing over intervals of Δ values for which the critical region for the level-α test of $H_{0,\Delta}$ against $H_{1,\Delta}$ does not change, but jumps (either up or down) whenever the critical region changes. Thus, finding Δ_α requires care. We used a three-step procedure that we describe now.

The first step in the procedure is to obtain an upper bound U for Δ_α by finding the value $U \equiv \Delta_{\alpha,IU}$ such that the p-value for the intersection–union test is α. The value $\Delta_{\alpha,IU}$ can be found by bisection since p_{IU} is a continuous, decreasing function of Δ. The second step in the procedure is to find a lower bound L for Δ_α by using bisection to find a value L (possibly not unique) such that the p-value either equals α at L or jumps past α at L. The third step is to compute p-values for a fine grid of Δ values that span the interval $[L, U]$ and take Δ_α to be the Δ value where the p-value for testing $H_{0,\Delta}$ using the improved test drops below α for good.

The second step in the procedure that we have just described involves finding a value L such that the p-value either equals α or jumps past α right at L. Such a point usually exists, but if the sample proportions $n_1/N, \ldots, n_k/N$ are all very close to $1/k$, it may happen that the p-value is less than α for every positive choice of Δ. In this case, we get $\Delta_\alpha = 0$, and the simultaneous confidence interval for $p_1, \ldots, p_k$ has length 0.

A simultaneous lower bound a_α for $p_1, \ldots, p_k$ can be obtained by inverting the family of tests of $H_0^- : \mathbf{p} \notin (a, 1]^k$ against $H_1^- : \mathbf{p} \in (a, 1]^k$, and a simultaneous upper bound b_α can be obtained by inverting the family of tests of $H_0^+ : \mathbf{p} \notin [0, b)^k$ against $H_1^+ : \mathbf{p} \in [0, b)^k$. Since the one-sided p-values used in doing these tests are strictly monotone functions of the parameter (a or b), either bound can be found using a root-finding algorithm like bisection. In cases where N is so large that the algorithm of Frey [11] is not sufficiently fast, conservative bounds $\Delta_{\alpha,IU}$, $a_{\alpha,IU}$, and $b_{\alpha,IU}$ can be computed in an obvious way by inverting the intersection–union test for the corresponding hypotheses. These bounds are often substantially looser than those obtained by inverting the improved test, but they are readily obtained even for extremely large N.

4. Power comparisons

To assess the power of the improved test, we compared it via simulation and direct calculation to that of possible alternative tests based on statistics from the power divergence family studied by Cressie and Read [4]. For $\eta \neq 0, -1$, the statistic I_η is defined by

$$I_\eta \equiv \frac{1}{N\eta(\eta+1)} \sum_{i=1}^{k} N_i \left\{ \left(\frac{N_i}{E_i} \right)^\eta - 1 \right\},$$

where $N_1, \ldots, N_k$ are the observed counts and $E_1, \ldots, E_k$ are the expected counts. The statistics I_0 and I_{-1} are determined by continuity so that, in particular, I_0 is the likelihood ratio test statistic $2\sum_{i=1}^{k} N_i \log(N_i/E_i)$. Since we seek evidence of closeness to $(1/k, \ldots, 1/k)$, we took the expected counts to be $E_i \equiv N/k$ for all i. In goodness-of-fit testing, one rejects when I_η is large. In this equivalence testing context, however, it makes sense to reject when I_η is small. The following theorem shows that critical values for the test of $H_0 : \mathbf{p} \notin (a, b)^k$ against $H_1 : \mathbf{p} \in (a, b)^k$ based on I_η may be obtained by simulating values of I_η at the points $\mathbf{p}_{a,k}$ and $\mathbf{p}_{b,k}$.

Theorem 2. *For $\mathbf{p} \notin (a, b)^k$, the distribution of I_η is stochastically smallest either when $\mathbf{p} = \mathbf{p}_{a,k}$ or when $\mathbf{p} = \mathbf{p}_{b,k}$.*

Proof. We present a proof here for the case where $\eta \neq 0, -1$. Similar arguments can be used for the $\eta = 0$ and $\eta = -1$ cases. First note that, by the same logic used in proving Theorem 1, it suffices to show that if $\mathbf{p} \equiv (p_1, \ldots, p_k)$ has $p_1 \neq p_2$, then moving p_1 and p_2 closer together (while keeping their sum constant) makes I_η stochastically smaller. As in the proof of Lemma 2, we use a conditioning argument.

Let $n_3, \ldots, n_k$ be given values for $N_3, \ldots, N_k$, and let $r > 0$ be arbitrary. Then the quantity $\frac{1}{N\eta(\eta+1)} \sum_{i=3}^{k} N_i\{(\frac{N_i}{E_i})^\eta - 1\}$ is given, and the conditional probability $P(I_\eta \leq r \mid N_3 = n_3, \ldots, N_k = n_k, \mathbf{p})$ can be rewritten as $P(\frac{N_1^{\eta+1}+N_2^{\eta+1}}{\eta(\eta+1)} \leq c \mid N_1 \sim \mathrm{Bin}(t, p_1/(p_1+p_2)))$ where $t = N - (n_3 + \cdots + n_k)$, c is an appropriate constant, and $N_2 = t - N_1$. Since the function $x \to \frac{x^{\eta+1}+(t-x)^{\eta+1}}{\eta(\eta+1)}$ is a convex function that is invariant under the transformation $x \to t - x$ and achieves a minimum at $t/2$, the list of N_1 values for which $\frac{N_1^{\eta+1}+N_2^{\eta+1}}{\eta(\eta+1)} \leq c$ is either empty or has the form $d, d+1, \ldots, t-d$ for some integer d. Thus, it follows from Lemma 1 that moving $p_1/(p_1+p_2)$ closer to $1/2$ either increases $P(I_\eta \leq r \mid N_3 = n_3, \ldots, N_k = n_k, \mathbf{p})$ or keeps it the same. Averaging the new conditional probabilities over the distribution of $(N_3, \ldots, N_k)$ as in the proof of Theorem 1 then completes the proof. □

We compared the power of the new test to that of tests based on I_η both by looking at the power on certain paths from the edges of the equivalence region to the center and by looking at the average power of each test over the entire equivalence region. We considered both linear paths from the point $\mathbf{p}_{b,k}$ to the center point $(1/k, \ldots, 1/k)$ and linear paths from the point $(b, b, (1-2b)/(k-2), \ldots, (1-2b)/(k-2))$ to the center point, and we also considered different choices of (a, b), N, k, and α. We initially considered a variety of different choices for η. However, we found that for η values in the range -2 to 6, the tests based on I_η were either similar in terms of power to the test based on I_2 or less powerful than the test based on I_2. Thus, in what follows, we present results only for the test based on I_2 and the improved test developed in Section 2. Some representative results from the power study are given in Figs. 1–4 and Table 2.

Figs. 1 and 2 show power results for the case where $k = 5$, $\alpha = 0.05$, $(a, b) = (0.1, 0.3)$, and $N = 100, 200, 400$, and 800. In each plot of Fig. 1, the probability vector $\mathbf{p}$ takes values $\mathbf{p}_\lambda = (1-\lambda)(0.3, 0.175, \ldots, 0.175) + \lambda(0.2, \ldots, 0.2)$, where $\lambda \in [0, 1]$. Thus, $\lambda = 0$ gives a point on the boundary of the equivalence region, and $\lambda = 1$ gives the center point $(0.2, \ldots, 0.2)$. The algorithm due to Frey [11] was used to compute exact power values for the improved test (solid curves), and simulated powers for the test based on I_2 (open dots) were obtained by doing 10000 runs at each λ value that was an integer multiple of 0.1. Critical values for the test based on I_2 were obtained via Theorem 2, with 100000 runs done at each of $\mathbf{p}_{a,k}$ and $\mathbf{p}_{b,k}$. We see from Fig. 1 that when $N = 100$, the test based on I_2 is more powerful than the new test. However, for larger N, the two tests are virtually the same in terms of power. In Fig. 2, the probability vector $\mathbf{p}$ takes values $\mathbf{p}_\lambda = (1-\lambda)(0.3, 0.3, 0.1\bar{3}, \ldots, 0.1\bar{3}) + \lambda(0.2, \ldots, 0.2)$, where $\lambda \in [0, 1]$, and we see that the test based on I_2 has an advantage for $N = 100$. For larger N, however, there is a clear advantage for the improved test, with the advantage being particularly large when $N = 800$.

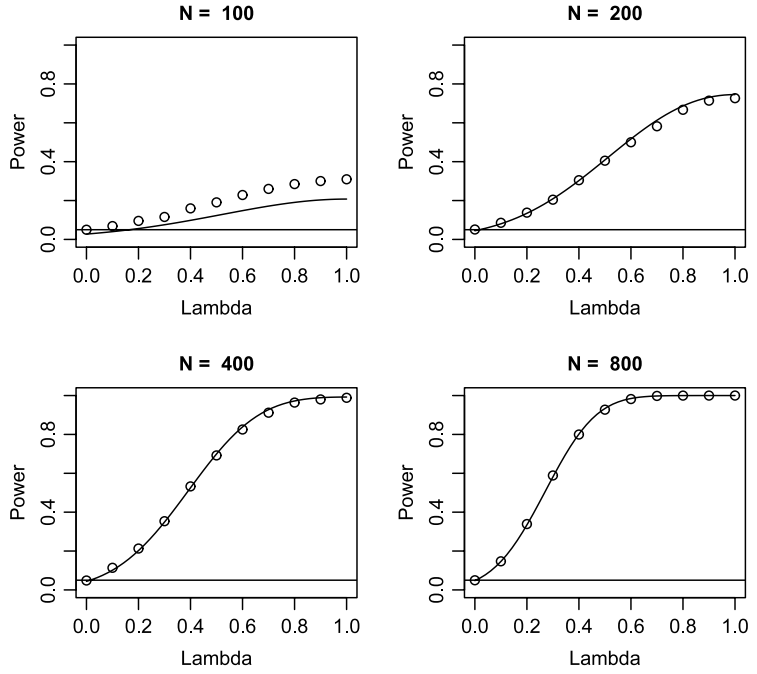

Fig. 1. Calculated power of the new test (solid line) and simulated power of the test based on I_2 (open dots) along the path from the point (0.30, 0.175, 0.175, 0.175, 0.175) to the point (0.2, 0.2, 0.2, 0.2, 0.2) when $k = 5, \alpha = 0.05$, $(a, b) = (0.1, 0.3)$, and N takes on the values 100, 200, 400, and 800. Each simulated power value was based on 10000 runs.

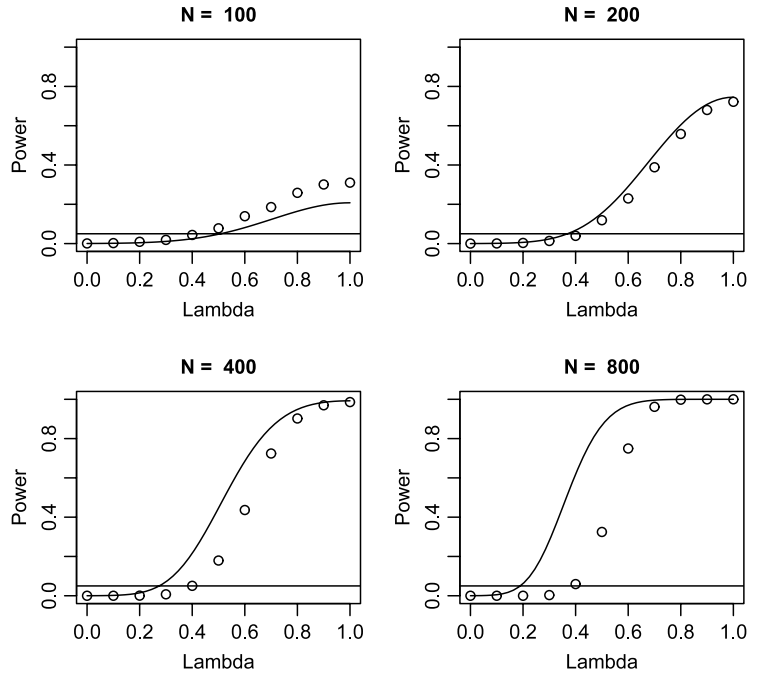

Fig. 2. Calculated power of the new test (solid line) and simulated power of the test based on I_2 (open dots) along the path from the point $(0.30, 0.30, 0.1\bar{3}, 0.1\bar{3}, 0.1\bar{3})$ to the point (0.2, 0.2, 0.2, 0.2, 0.2) when $k = 5, \alpha = 0.05$, $(a, b) = (0.1, 0.3)$, and N takes on the values 100, 200, 400, and 800. Each simulated power value was based on 10000 runs.

Figs. 3 and 4 are the analogs of Figs. 1 and 2 for the case where $k = 10, \alpha = 0.05$, and $(a, b) = (0.05, 0.15)$. We see from the two figures that, as in Figs. 1 and 2, the test based on I_2 has an advantage in terms of power when $N = 100$. However, for larger values of N, the new test is either just as powerful (for $N = 200, 400$) or noticeably more powerful ($N = 800$)

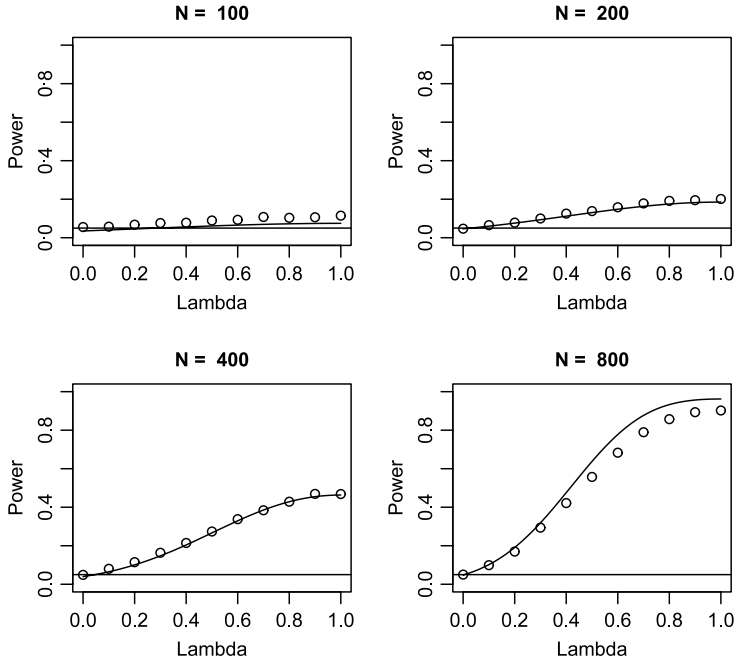

Fig. 3. Calculated power of the new test (solid line) and simulated power of the test based on I_2 (open dots) along the path from the point $(0.15, 0.09\bar{4}, \ldots, 0.09\bar{4})$ to the point $(0.1, \ldots, 0.1)$ when $k = 10$, $\alpha = 0.05$, $(a, b) = (0.05, 0.15)$, and N takes on the values 100, 200, 400, and 800. Each simulated power value was based on 10000 runs.

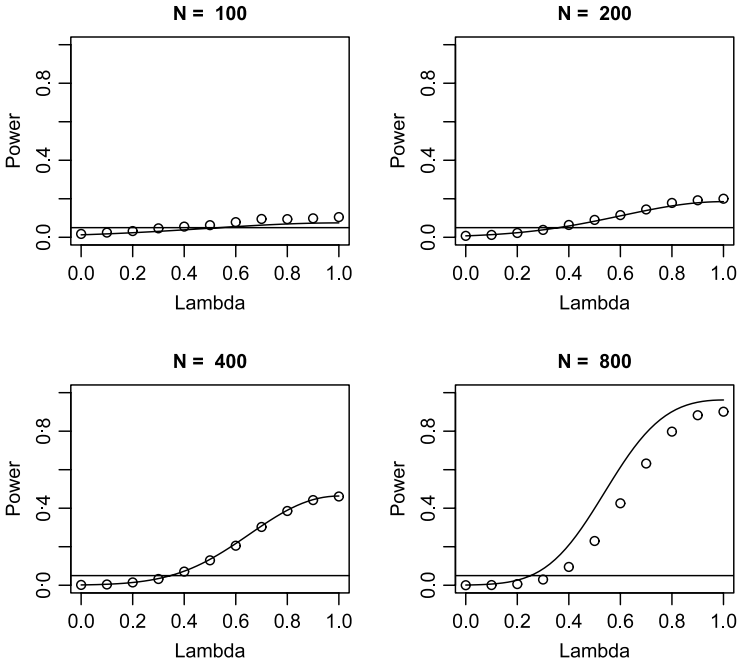

Fig. 4. Calculated power of the new test (solid line) and simulated power of the test based on I_2 (open dots) along the path from the point $(0.15, 0.15, 0.0875, \ldots, 0875)$ to the point $(0.1, \ldots, 0.1)$ when $k = 10$, $\alpha = 0.05$, $(a, b) = (0.05, 0.15)$, and N takes on the values 100, 200, 400, and 800. Each simulated power value was based on 10000 runs.

than the test based on I_2. Overall, the four figures suggest that while the test based on I_2 has an advantage when N is small, the new test is the more powerful of the two for larger N values. In fact, the advantage for I_2 seems to be restricted to cases where neither test has good power. One reason for the advantage of the test based on I_2 when N is small is that the size of the improved test is often much less than the α level when N is small. With larger N, there are more potential critical regions, and the size of the improved test tends to be closer to α.

Table 2
Simulated average power over the entire equivalence region for the new test and the test based on I_2 when $k = 5$ and $(a, b) = (0.1, 0.3)$. Each simulated power value was based on 10000 runs.

$\alpha = 0.05$			$\alpha = 0.10$		
N	I_2	New	N	I_2	New
100	.0344	.0195	100	.0790	.0589
200	.0541	.0809	200	.1001	.1429
400	.0930	.1895	400	.1412	.2737
800	.1485	.3440	800	.1857	.4199

The power plots in Figs. 1–4 were based on probability vectors chosen from just a few line segments in a high-dimensional equivalence region. To make a more comprehensive power comparison, we examined the average power of each test over the entire equivalence region. We took the Dirichlet distribution with parameter $(1, \ldots, 1)$ to be the uniform distribution on the entire parameter space, and for each choice of equivalence region, we generated 10000 probability vectors from the restriction of this distribution to the equivalence region. For each vector $\mathbf{p} \in (a, b)^k$ that we generated, we generated one data point $(N_1, \ldots, N_k) \sim \text{Multi}(N, \mathbf{p})$ and conducted both tests. Each test's average power was then estimated as the proportion of the 10000 tests for which $H_0 : \mathbf{p} \notin (a, b)^k$ was rejected. Average power results for the case where $k = 5$ and $(a, b) = (0.1, 0.3)$ are given in Table 2, which gives results both for $\alpha = 0.05$ and for $\alpha = 0.10$. We see from the table that when $N = 100$, the test based on I_2 has better average power. However, for larger N, the new test has far better average power. Indeed, for $N = 800$, the average power for the new test is more than double that of the test based on I_2. Thus, taking all the comparisons into account, the power properties of the new test compare favorably to those of the test based on I_2.

5. Examples

To illustrate the new test and the associated confidence interval procedure, we applied them to multiple data sets where concluding that the outcomes are approximately equally likely would be of interest. We first tested for evidence that the digits 0 to 9 are approximately equally likely to appear as digits of π. We then tested for evidence that certain roulette wheels gave approximately equally likely outcomes. Finally, we tested for positive evidence of fairness in the selection of digits 0 to 9 in the mid-day Big 4 drawing run by the Pennsylvania Lottery. The first and third examples are cases where the outcomes might actually be equally likely, while the second example is one where the outcomes are probably not equally likely, but may be approximately equally likely.

The data on the digits of π were counts, obtained from [13], of the number of times each digit 0 to 9 appears among the first 10^r digits of π for $r = 2, \ldots, 6$. These counts are given in Table 3, and we treated them as summaries of independent draws from a distribution on the numbers 0 to 9. Testing whether the digits of π are actually independent is another interesting problem (see [12]), but we focus here on obtaining evidence that the digits are approximately equally likely to appear. Using the improved test or (for cases where N is very large) the intersection–union test, we computed the upper bound $\Delta_{0.05}$ for each of the data sets. We also, for comparison purposes, report the results of a chi-squared goodness-of-fit test of $H_0 : \mathbf{p} = (0.1, \ldots, 0.1)$ against $H_1 : \mathbf{p} \neq (0.1, \ldots, 0.1)$. The chi-squared test does not lead to rejection of the homogeneity null hypothesis at level 0.05 for any of the five data sets, but the equivalence test allows us to find increasingly strong evidence that the outcomes are approximately equally likely as N increases. With $N = 100$, we are able to conclude only that each of the outcome probabilities falls in the interval (0.0030, 0.1970) (since $\Delta_{.05} = 0.0970$). However, with $N = 1000000$, we are able to conclude that each of the outcome probabilities falls in the narrow interval (0.0991, 0.1009). The starred upper bounds in Table 3 and in the remainder of this section were obtained by inverting the intersection–union test.

Roulette data were obtained both from [5], who collected data to investigate whether the betting behavior of gamblers is consistent with belief in the gambler's fallacy and the hot hand, and from [20], who collected data in hope of finding a biased wheel and making money. Key data on one wheel from [5] and two wheels from [20] are given in Table 4. Both of the Wilson [20] wheels were located at Harold's club in Reno, NV, and the data are given on pages 292–295 of [20]. A double-zero roulette wheel has 38 different compartments labeled 00, 0, 1, ..., 36 that are intended to be equally likely, and one can bet either on a single number (which pays \$35 on a \$1 bet) or on certain combinations of numbers. In addition to providing the total number of runs and the minimum and maximum counts over the 38 outcomes, Table 4 also gives the value $\Delta_{.05}$ and the upper bound $b_{.05}$ obtained by inverting the exact test of $H_0^+ : \mathbf{p} \notin [0, b)^k$ against $H_1^+ : \mathbf{p} \in [0, b)^k$. We see that for the Croson and Sundali [5] wheel, the chi-squared test does not find sufficient evidence to reject the hypothesis of equally likely outcomes. However, the relatively large bound $\Delta_{0.05} = 0.0182$ indicates that there is also little evidence that the outcomes are approximately equally likely. For the two Wilson [20] wheels, in contrast, the chi-squared test allows one to conclude that the outcomes are not equally likely, but the equivalence test provides some positive evidence that the outcomes are approximately equally likely. For example, the simultaneous confidence interval for the outcome probabilities for Wilson's moderately biased wheel is (0.0230, 0.0296).

Since a successful bet on a single number pays \$35 on a \$1 bet, a casino operator can ensure that there are no money-losing numbers by confirming that each p_i is no larger than 1/36. This suggests testing $H_0^+ : \mathbf{p} \notin [0, 1/36)^{38}$ against

Table 3
Data on the digits of π and associated summaries. The starred $\Delta_{.05}$ values were obtained by inverting the conservative intersection–union test.

Digit	$N = 10^2$	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$
0	8	93	968	9999	99959
1	8	116	1026	10137	99758
2	12	103	1021	9908	100026
3	11	102	974	10025	100229
4	10	93	1012	9971	100230
5	8	97	1046	10026	100359
6	9	94	1021	10029	99548
7	8	95	970	10025	99800
8	12	101	948	9978	99985
9	14	106	1014	9902	100106
χ^2	4.20	4.74	9.32	4.09	5.51
p-value	0.8978	0.8564	0.4085	0.9052	0.7879
$\Delta_{.05}$	0.0970	0.0312	0.0087	0.0030*	0.0009*

Table 4
Key summary data from various 38-compartment roulette wheels. The starred bounds were obtained by inverting the conservative intersection–union test.

Wheel	N	(n_{min}, n_{max})	$\Delta_{0.05}$	$b_{.05}$	χ^2 (p-value)
Croson and Sundali [5]	904	(15, 33)	0.0182	0.0446	31.2 (0.7370)
Wilson [20] strongly biased	11766	(271, 385)	0.0092*	0.0355*	78.8 (0.0001)
Wilson [20] moderately biased	79800	(1912, 2284)	0.0033*	0.0296*	117.0 (0.0000)

$H_1^+ : \mathbf{p} \in [0, 1/36)^{38}$. We see from Table 4 that for these wheels the upper bounds $b_{.05}$ all exceed $1/36 = .02\bar{7}$. Thus, one is unable to conclude at level 0.05 that none of the numbers offer an advantage to the gambler.

The Pennsylvania Lottery Big 4 game depends on a random 4-digit number created from four separate drawings of a single digit (with repeats allowed). Winning numbers for the daily mid-day drawings in 2008 and 2009 were obtained from [14]. Since two winning numbers were listed for March 24, 2008, the total number of digits was 1468 for 2008 (due to the leap year and the extra number) and 1460 for 2009. The data are given in Table 5. The chi-squared goodness-of-fit test indicates that there is no reason to doubt the hypothesis of equally likely outcomes in either year, while the equivalence test gives simultaneous 95% confidence intervals of (0.0690, 0.1310) (for 2008) and (0.0725,0.1275) (for 2009).

The simultaneous confidence intervals obtained here are comparable in length to those we would have obtained using other methods in the literature. For example, Fitzgerald and Scott [9] showed that using intervals of the form $\hat{p}_i \pm 1.13/\sqrt{N}$, where $\hat{p}_i = N_i/N$, gives an asymptotic simultaneous coverage probability of at least 95%. Using such intervals here would lead to margins of error of 0.0295 (for 2008) and 0.0296 (for 2009). These are comparable to the margins of error 0.0310 and 0.0275 that we obtained from inverting the equivalence test. However, since the Fitzgerald and Scott [9] intervals are each centered at $\hat{p}_i$ rather than at 0.1, they are less helpful for drawing strong conclusions about the maximum or minimum of the outcome probabilities.

6. An alternate equivalence region

The improved test developed in Section 2 is not an unbiased test. In fact, we saw in Section 4 that there are points inside the equivalence region at which the test has virtually no power at all. Modifying the critical region is one strategy for reducing the amount of bias. Another strategy is to modify the equivalence region. Specifically, rather than using an equivalence region of the form $(a, b)^k$ for a and b satisfying $0 \le a < 1/k < b \le 1$, we might choose the equivalence region to be the set of all points where the test that rejects when $n_i \in [c_l, c_u]$ for $i = 1, \ldots, k$ has power α or higher. Multiple applications of Lemma 2 show that such an equivalence region is star-like with respect to the center point $(1/k, \ldots, 1/k)$, and if $k = 3$, it is possible to make pictures of such regions.

Fig. 5, plotted in barycentric coordinates, shows the equivalence region (the hexagon) and some power contours of the new equivalence test when $N = 85$, $\alpha = 0.05$, $k = 3$, and $(a, b) = (1/3 - 1/5, 1/3 + 1/5)$. The contours are for powers of 0.05, 0.20, 0.40, 0.60, 0.80, and 0.90. We see that while the 0.05 power contour nearly touches each side of the hexagon, there are areas at the corners of the equivalence region where the power is less than 0.05. However, if we choose the equivalence region to be the interior of the 0.05 power contour, then the test becomes unbiased. This is an attractive option for $k = 3$, but would be difficult to visualize in higher dimensions.

7. Conclusions

We have developed a new exact equivalence test that allows one to conclude that k possible outcomes are approximately equally likely. This test compares favorably to possible alternative tests in terms of power, and it may be inverted to yield

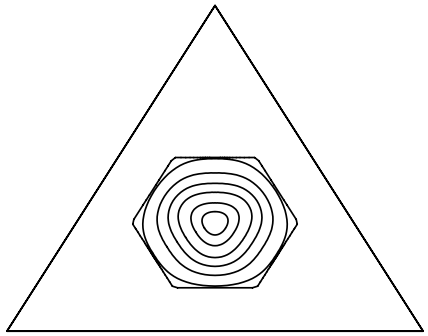

Fig. 5. The equivalence region when $N = 85, k = 3$ and $(a, b) = (1/3 - 1/5, 1/3 + 1/5)$, together with power contours for the new level-0.05 equivalence test. The contours are for powers of 0.05, 0.20, 0.40, 0.60, 0.80, and 0.90, and the points are plotted in barycentric coordinates so that the vertices of the triangle represent the probability vectors (1, 0, 0), (0, 1, 0), and (0, 0, 1).

Table 5
Number of times each digit 0 to 9 appeared as a winning number in the mid-day big 4 drawings of the Pennsylvania Lottery in 2008 and 2009.

Digit	2008	2009
0	156	154
1	156	154
2	164	153
3	132	148
4	156	138
5	135	151
6	117	155
7	161	138
8	135	148
9	156	121
χ^2 statistic	15.13	7.15
p-value	0.087	0.622
$\Delta_{0.05}$	0.0310	0.0275

two-sided or one-sided simultaneous confidence intervals for the outcome probabilities $p_1, \ldots, p_k$. We have applied the test to several different data sets where establishing that outcomes are approximately equally likely is of interest. In doing so, we have shown that the conclusions obtained using the equivalence test are different, and often more relevant, than those obtained from a goodness-of-fit test.

Acknowledgments

The author thanks the referees for helpful suggestions that have improved the paper.

References

[1] K.E. Atkinson, An Introduction to Numerical Analysis, second edition, Wiley, New York, 1989.
[2] R.L. Berger, Multiparameter hypothesis testing and acceptance sampling, Technometrics 24 (1982) 295–300.
[3] W.H. Bondy, A test of an experimental hypothesis of negligible difference between means, The American Statistician 23 (5) (1969) 28–30.
[4] N. Cressie, T.R.C. Read, Multinomial goodness-of-fit tests, Journal of the Royal Statistical Society, Series B 46 (1984) 440–464.
[5] R. Croson, J. Sundali, The gambler's fallacy and the hot hand: empirical data from casinos, Journal of Risk and Uncertainty 30 (2005) 195–209.
[6] H.A. David, H.N. Nagaraja, Order Statistics, third edition, Wiley, New York, 2003.
[7] P.M. Dixon, J.H.K. Pechmann, A statistical test to show negligible trend, Ecology 86 (2005) 1751–1756.
[8] S.N. Ethier, Testing for favorable numbers on a roulette wheel, Journal of the American Statistical Association 77 (1982) 660–665.
[9] S. Fitzgerald, A. Scott, Quick simultaneous confidence intervals for multinomial proportions, Journal of the American Statistical Association 82 (1987) 875–878.
[10] J. Frey, An exact multinomial test for equivalence, Canadian Journal of Statistics 37 (2009) 47–59.
[11] J. Frey, An algorithm for computing rectangular multinomial probabilities, Journal of Statistical Computation and Simulation 79 (2009) 1483–1489.
[12] T. Jaditz, Are the digits of π an independent and identically distributed sequence? The American Statistician 54 (2000) 12–16.
[13] Kanada Laboratory. Sample digits for decimal digits of π, 2010, Available at: http://www.super-computing.org (accessed 28.09.10).
[14] Pennsylvania Lottery. Past winning numbers, 2010, Available at: http://www.palottery.state.pa.us/past-winning-numbers.aspx (accessed 28.09.10).
[15] E.L. Lehmann, Testing Statistical Hypotheses, Wiley, New York, 1959.
[16] A.P. Robinson, R.E. Froese, Model validation using equivalence tests, Ecological Modelling 176 (2004) 349–358.
[17] N.H. Spencer, Overcoming the multiple-testing problem when testing randomness, Applied Statistics 58 (2009) 543–553.
[18] S. Wellek, Testing Statistical Hypotheses of Equivalence, Chapman and Hall, New York, 2003.
[19] W.J. Westlake, Symmetric confidence intervals for bioequivalence trials, Biometrics 32 (1976) 741–744.
[20] A.N. Wilson, The Casino Gambler's Guide, Harper and Row, New York, 1965.